SYSTEM FOR DETERMINING DEGREES OF SIMILARITY IN EMAIL MESSAGE INFORMATION

ABSTRACT OF THE DISCLOSURE

Similarity of email message characteristics is used to detect bulk and spam email. A determination of "sameness" for purposes of both bulk and spam classifications can use any number and type of evaluation modules. Each module can include one or more rules, tests, processes, algorithms, or other functionality. For example, one type of module may be a word count of email message text. Another module can use a weighting factor based on groups of multiple words and their perceived meanings. In general, any type of module that performs a similarity analysis can be used. A preferred embodiment of the invention uses statistical analysis, such as Bayesian analysis, to measure the performance of different modules against a known standard, such as human manual matching. Modules that are performing worse than other modules can be valued less than modules having better performance. In this manner, a high degree of reliability can be achieved. To improve performance, if a message is determined to be the same as a previous message, the previous computations and results for that previous message can be re-used. Users can be provided with options to customize or regulate bulk and spam classification and subsequent actions on how to handle the classified email messages.